The One About Ethics

Transcript of the Human-Robot Interaction Podcast Episode 32. Published on 1 November 2024 by Christoph Bartneck at:

https://www.human-robot-interaction.org/2024/11/01/the-one-about-ethics/

[00:00:00] **Christoph Bartneck:** Robots are not just machines. We treat them as if they were somewhat like humans, including applying moral standards and expectations to them. Our behavior towards robots matters. The way we treat them reveals much about ourselves. In today's episode, we will talk about how being nice is not the opposite of being cruel to them. I invited Bob Douglas and Mary Blossom from the AI Research Institute to introduce us to the topic. They agreed to produce a short podcast dialogue to get us started.

Bob and Mary's Introduction

[00:00:56] **Bob:** Okay. So like ever noticed how much broken zipper just like really bugs you way more than a zipper that works perfectly. It makes you happy.

[00:01:07] **Mary:** It's like

[00:01:07] **Bob:** our brains just zero in on the bad stuff, you know, and get this, it's not just like a human quirk, turns out we do the same thing with robots.

[00:01:16] **Mary:** Really?

That's wild.

[00:01:18] **Bob:** Yeah, we're kind of brutal with our robot judgments, like way harsher on their mess ups than, I don't know, impressed by when they do something right.

[00:01:26] **Mary:** So it's like, what, we're holding robots to a way higher standard than we hold ourselves?

[00:01:31] **Bob:** That's exactly what we are diving into today. We've got a stack of research papers all about it, the morality of humans and robots.

Who knew that was even a thing, right?

[00:01:38] **Mary:** Well, it's more common than you think. And it makes sense, right? Like, we talk about robots being the future and all, but we don't always stop and think. Yeah. How does our sense of right and wrong even A P P L Y to a robot brain?

[00:01:52] **Bob:** Yeah. It's like we're giving them grades, but the Fs are in, like, Hugh G E bold font.

[00:01:58] **Mary:** That's a great way to put it. And that's essentially what a lot of these researchers are finding. They call it asymmetric morality, basically. Yeah. A robot's bad deeds, they just carry way more weight for us than their good deeds, even if the actions themselves are pretty similar in impact.

[00:02:11] **Bob:** Like a robot could save a thousand puppies.

[00:02:14] **Mary:** Right. A thousand puppies.

[00:02:16] **Bob:** Steps on one toe.

[00:02:17] **Mary:** Suddenly it's a menace.

[00:02:19] **Bob:** Exactly. So like, what's the deal with that? Why are we so quick to judge?

[00:02:22] **Mary:** Well, there are a few different theories. One that I find particularly interesting is this idea of the practical wisdom problem.

[00:02:29] **Bob:** The what now? Okay, break that down for me.

What does practical wisdom even have to do with robots?

[00:02:35] **Mary:** So think of it this way, as humans, part of being, you know, a good human, it's about being virtuous, making moral choices.

[00:02:42] **Bob:** Right. Trying to do the right thing most of the time.

[00:02:44] **Mary:** Exactly. True virtue, it means you understand the context of your actions, the nuances, the why behind what you're doing, and robots Well, let's just say they're not exactly known for their grasp of human nuance.

[00:03:00] **Bob:** Okay, so like, if my Roomba cleans the whole floor and I'm all, Good robot! You did amazing! It's not like the Roomba knows it did a good job. It was just following its programming.

[00:03:08] **Mary:** Exactly. It's like praising a calculator for spitting out the right answer. That praise, while, you know, probably well intentioned, it's kind of misplaced.

[00:03:16] **Bob:** It's like, The robot didn't earn that moral praise the same way another person would.

[00:03:20] **Mary:** Exactly. There was actually this study where people were less likely to trust a robot that kept apologizing for a mistake because it just felt off, like too much, almost manipulative, even though it obviously wasn't intentional on the robot's part.

And then on top of that, you've got the whole negativity bias thing.

[00:03:35] **Bob:** Which is, don't leave me hanging.

[00:03:37] **Mary:** Oh, right. Sorry. So hardwired to focus on threats, right? Like, it's an evolutionary leftover from back when we had to constantly dodge saber toothed tigers and whatnot. Makes

[00:03:47] **Bob:** sense. Better safe than sorry, right?

[00:03:49] **Mary:** Exactly. So, when a robot screws up, even if it's something small, it disrupts our expectations of how machines should act. And that sets off those alarm bells in our heads. It's like, hold on, this thing isn't acting the way it's supposed to.

[00:04:01] **Bob:** Even a little glitch. can make us question a robot's whole deal.

Like, is this the start of the robot uprising? Should I be worried? So, a robot could, like, bake me a cake perfectly iced and everything. With,

[00:04:12] **Mary:** like, little robot sprinkle.

[00:04:14] **Bob:** Exactly. But if it trips on the way to the table, and, I don't know, spills a crumb.

[00:04:20] **Mary:** It's ruined forever. Yeah, no, I totally get that. It's not always about, like, awesome versus awful, though, right?

It's more like, does the robot's action, good or bad, does it make sense for how we expect that kind of machine to act?

[00:04:33] **Bob:** Okay, so, like, if my coffee maker, right, if it, like, suddenly started doing a stand up routine, that'd be more alarming than if it just, like, Remember to use oat milk.

[00:04:41] **Mary:** Exactly. It violates our mental model of what a coffee maker should be.

And that actually brings up another important point about all this. It's like when a robot does something bad, we're more likely to think, uh oh, this is just the beginning. Next thing you know, it's going to be taken over the world. But if it does something good, we're like, yeah, well it was programmed that way.

Like it doesn't get the same credit.

[00:05:02] **Bob:** Right. Right. Like it has to prove itself way more. Okay. So we've talked about like why we judge robots this way. Yeah. Yeah. Boo. What does this actually mean? Like, out in the real world, does this matter?

[00:05:12] **Mary:** Oh, absolutely it matters. Like, big time. Think about, for instance, how people are actually designing robots now.

It's not enough to just program it to, I don't know, sort widgets or whatever. Right. They've got to factor in this whole judgment gap thing. They have to think about how even tiny errors, like things that a human might brush off, could totally erode someone's trust in the whole robot.

[00:05:34] **Bob:** So, like, if a robot's a little too eager to please It could backfire, even if it's trying to be helpful.

[00:05:41] **Mary:** Exactly. Like, imagine a robot assistant that keeps offering you coffee refills, even after you've said no thanks, like, five times.

[00:05:48] **Bob:** Oh my gosh, the worst. The worst.

[00:05:50] **Mary:** Right. It might think it's being helpful, but to us humans, it's just not. Intrusive. And you're probably going to trust that robot less after that, right?

[00:05:57] **Bob:** For sure. Like, read the room, robot. Okay, but that brings up something else, too. Like, what happens when a robot actually does cause harm? Are we more likely to want to punish it, like deactivate it or something? Versus, are we as quick to reward them when they do good?

[00:06:13] **Mary:** That's the million dollar question.

It's almost like, are we gonna have robot jail, robot rewards programs? And honestly, there are no easy answers.

[00:06:21] **Bob:** So it's not so simple as good robot or bad robot.

[00:06:24] **Mary:** Right. One study found that, yeah, maybe we don't fully blame a robot if it messes up, especially if it seems like an accident. Yeah. But we still want someone to fix it so it doesn't happen again.

[00:06:34] **Bob:** So it's like, even if we don't think a robot is morally responsible like a human, We still expect consequences. Like with that self driving car example, if it swerves to avoid hitting, like, a squirrel, and ends up denting your car, do you blame the car?

[00:06:50] **Mary:** Right, right, exactly. It's tricky.

[00:06:52] **Bob:** Maybe not. But do you want someone to, like, reprogram that car so it doesn't happen again?

Probably.

[00:06:56] **Mary:** Yes, a hundred percent. This is showing us that how we feel about robots, our moral intuitions, they're totally evolving at the same time the tech itself is evolving, which is kind of wild when you think about it. It's

[00:07:10] **Bob:** like we're making up the rulebook for robot morality as we go. Which is kind of a lot of pressure, if you ask me.

It's almost like we're trying to, like, shove this robot shaped block into our human morality hole.

[00:07:22] **Mary:** And that's what I think makes this whole thing so fascinating. It's not even really about programming the robots, is it? It's about us. Like, we're the ones with all these ingrained ideas about good and bad.

And we just, like, expect the robots to follow along. And it's

[00:07:35] **Bob:** not even like we're perfect at it ourselves. We all mess up, right?

[00:07:37] **Mary:** Exactly. But

[00:07:38] **Bob:** there's something about robots where we expect I don't know, more.

[00:07:42] **Mary:** It's like we're projecting, right? Like all our hopes and fears onto these machines. So on one hand, be perfect, be helpful, be super efficient.

But on the other hand, don't get any ideas, right? Don't go all Terminator on us.

[00:07:55] **Bob:** Yeah, those sci fi stories really did something to us, huh? Even if it's just, like, a Roomba bumping into a wall too many times.

[00:08:02] **Mary:** Totally. So even as they get more and more advanced, better at acting human, That little knot of anxiety, I don't think it's going away anytime soon.

[00:08:11] **Bob:** So what now? Are we stuck like this forever? Judging robots way harsher than we judge ourselves?

[00:08:17] **Mary:** It's a big question. But the research, I think it gives us a starting point. At least now we can be aware of it, right? Like, next time you see a robot doing its thing, good or bad, just take a second. Ask yourself, why am I reacting this way?

Am I being fair?

[00:08:29] **Bob:** It's like, Robot Morality 101 is actually about us examining our own moral compasses.

[00:08:36] **Mary:** Because at the end of the day, the future of all this, it's not just about tech. It's how we deal with this stuff inside ourselves.

[00:08:43] **Bob:** Bridging that gap between the robots we build and the values we, you know, try to live by.

[00:08:49] **Mary:** Exactly. And hey, maybe if we figure out how we judge them, we might learn a thing or two about how we judge each other, too.

[00:08:55] **Bob:** Which I mean, wouldn't be the worst thing in the world, right? Thanks for diving in with us today, everybody. Be sure to check out the show notes for links to all the research we talked about and until next time, keep those robot morality questions coming.

Interview

[00:09:20] **Christoph Bartneck:** Bob and Mary are not real. They are a product of Google's Notebook LM. For those of you who noticed, kudos. For you who didn't, I'm sorry to have misled you. In Notebook LM, you can upload several documents and it creates this type of podcast for you. I wonder if the Human Robot Interaction podcast can soon be automated. To discuss this, I invited Michael-John Turp and Mingyi Wang from the University of Canterbury to the table. Minyi, could you please introduce yourself?

[00:09:53] **Minyi:** So, hi everyone. I'm Minyi Wang, and I'm a PhD student in University of Canterbury.

[00:10:00] **Christoph Bartneck:** And Michael, what are you doing here?

[00:10:01] **Michael-John:** That's a good question. Well, so my name is Michael John Turp. I'm a lecturer in philosophy, and I specialize in moral psychology and meta ethics here at the University of Canterbury.

[00:10:10] **Christoph Bartneck:** Wonderful. And I brought you in today because I sent you this. Podcast Recording that was done by Google's notebook LM and I wonder what do you think about it? So shall we start maybe with Michael? Is there anything that was missing from the podcast?

[00:10:32] **Michael-John:** Well, so it was a very high level overview of some papers that were fed into it And of course there wasn't a great deal of concrete detail and I say overall I was very impressed with the tone. It's very impressive how these large language models can generate text. And it seemed, on its face, pretty persuasive. But then when you look down into the details, you see that there is quite a lot missing.

[00:10:56] **Christoph Bartneck:** Okay, and what are examples of things that are missing?

[00:10:59] **Michael-John:** Well, it's dealing with a range of papers at a very high level. And I think this is typical of LLMs, things like chat GPT as well, that they will give you this bird's eye view. And of course some of the examples that they give might be rather tangentially related to the actual context of the content of the papers.

[00:11:16] **Christoph Bartneck:** Minyi, what do you think? Was there anything missing from the podcast?

[00:11:19] **Minyi:** Actually the first time when I see the podcast I think, Oh, that's really amazing about that. But when I look through and review it again and again and again, I think there's something missing in that. So the first thing that I would like to talk about is about the framework. Because for my doctoral research is to explore the ethical asymmetry in human robot interaction.

Specifically this hypothesis came up with Robert Sparrow in 2021, and he highlighted a potential asymmetry in how human evaluate virtuous and vicious behaviors towards robot based on the virtual ethics. But for the podcast, the virtue ethics doesn't be mentioned. And in my opinion when it comes to conducting a research or exploring a phenomena the conceptual framework is the most significant factor in making the research more rational and getting how to conduct the research. So in my research virtue ethics is primary framework.

[00:12:22] **Michael-John:** Mean, a nice example is the way it starts with the zipper example, which I don't think is there in any of the papers. And, of course, doesn't apply specifically to human robot interaction. A zipper is, of course, an old technology. It's not a robot. It's a very engaging example to get you into it, but in another way, it can kind of misframe the whole thing.

[00:12:42] **Minyi:** Speaking of the example they provided I would like to say the zipper example is a little bit weird for me, because it's related to a negativity bias, and it's really missed in the study.

our papers, I think. I reviewed all the papers we uploaded, but no one mentioned the terms of negativity bias. I think it's generated automatically by AI. And for this negative negativity bias I don't think, I think that everyone will be upsided by the broken things right. And I don't think this is the relative to negative bias. It's a little bit weird for me.

[00:13:21] **Christoph Bartneck:** Well, in all fairness, we didn't ask it to summarize your thesis, we just asked it, here are six papers, summarize them. So did all of these papers have virtue ethics in it, or was that just one of many frameworks mentioned in the papers?

[00:13:42] **Minyi:** I think some of the papers mentioned virtue ethics especially in, I think we uploaded the Sparrows papers, which is the basic papers for my research.

Some of them mention the virtue ethics, but I think the podcast, the LLM doesn't catch this.

[00:14:01] **Christoph Bartneck:** So it's a bit lacking in detail?

[00:14:03] **Minyi:** You mean yeah for me, the I think the virtue ethics is most important. But interestingly, I will also mention that a podcaster mentioned that practical wisdom. This is one of master virtue in the virtue ethics. But obviously it didn't explain good for me. All

[00:14:22] **Christoph Bartneck:** So can we maybe summarize that it did bring interestingly some new content to the table? While neglecting a lot of detail. Is that a fair summary of the podcast?

[00:14:35] **Michael-John:** I mean, yes, I'd say that's a fair summary. And I think that's quite typical of how LLMs and LMs in this case do operate that sometimes you have rather loose associations. A couple of the examples I thought were rather fun for example, the example with the coffee and another example with the dog saving a thousand puppies as opposed to standing on a someone's toe.

If you look in the original papers, then you do see, interestingly, references both to coffee and to dogs, but in quite different contexts and to make quite different points. So, for example, the idea of coffee was mentioned as one kind of norm, that some people will have milk with their coffee. Dogs were mentioned in the context of Kantian ethics.

So I imagine that the the language models picked up on these terms and use them to generate examples, which in these cases don't exactly map onto the original papers and the arguments being made.

[00:15:33] **Christoph Bartneck:** So let me be more critical here. Did the model make any mistakes?

[00:15:40] **Minyi:** True, true. That speaking of the example of the robot saving the puppies and the step on one toes. The podcaster used this example to raise the practical wisdom in the moral judgment. That's really funny. I would like to explain the practical wisdom first because it is accurate appreciation of situation and the practical knowledge regarding how to deal with it. So facing this two situation, if people think robot become a threat when the robot steps on one toe after saving A thousand, thousand of puppies,I don't think that people have the good practical wisdom.

[00:16:23] **Christoph Bartneck:** So that's a clear mistake?

[00:16:24] **Minyi:** Yeah that's a mistake, I think. I

[00:16:26] **Michael-John:** Certainly it's an example which doesn't illustrate very well the basic point that is being made in the papers.

[00:16:32] **Christoph Bartneck:** So, to remind our audience, what was the exact situation that the language model described?

[00:16:37] **Michael-John:** So the situation was imagine that you have a robot, and in one situation, it saves a thousand puppies, which is, you might think, a very heroic sort of thing to, to do. In another situation, it just steps on a person's toe, which you might think is rather A minor misdeed. And according to the content of the podcast we would condemn harshly the robot for this single minor misdeed and not praise it for saving a thousand puppies.

Now, of course, that does illustrate this broader idea of asymmetry, but in this case, you have such a difference between the good outcome and the bad outcome that it doesn't really reflect the underlying research. very much.

[00:17:16] **Christoph Bartneck:** So that is wrong.

[00:17:18] **Michael-John:** Well, it's misleading. Let's say at the least. I mean, so it's using an example which it has produced which has, you know, certain structural similarities, let's say with the examples in the research. But if you are just exposed to the content of the podcast, I think it would be possible to come away with quite a misleading impression. I mean,

[00:17:38] **Christoph Bartneck:** by that logic, any robot doing the slightest mistake would be quite severely frowned upon or even punished, whereas saving puppies is a trivial thing. So

[00:17:52] **Michael-John:** that's right, and of course it raises the issue of should we have robot jails and so on and seems to be pretty agnostic on this question, whereas I think most reasonable people would conclude that robot jails are not the way forward.

[00:18:05] **Christoph Bartneck:** But are there any other mistakes more along the line of the theory in terms of the arguments that are being presented about something being correct or incorrect.

[00:18:16] **Michael-John:** So I think on the whole it reflects quite well the the basic ideas of the theories. But it does so very quickly and superficially and again using misleading examples. So, if you're familiar both with the papers that it's based on and the podcast, then it's much easier to navigate your way through that.

If on the other hand, I imagined myself in the position of someone who only has the podcast to go on, then I might arrive at, you know, rather distorted views of the underlying theory, especially the kind of enormous asymmetries that we were talking about between, let's say, saving a thousand puppies lives as opposed to, you know, standing on someone's toe.

[00:18:57] **Minyi:** Okay I would like to mention the negative bias again. It seems correct because previous research shows that the negative events can bring a greater psychological effect on attention or learning memory compared to the positive events. All the research that in the paper Or in the psychology field. But is it the same in the human robot interaction field? We don't know that. But, the podcast shows that it seems can be the same in the HRI field. I think that's one of the

[00:19:34] **Michael-John:** Indeed, so it fails to draw the kind of distinctions that we as academics want to draw and that's a bread and butter. Another thing I was actually mentioning to you before Christoph is that I found when listening to it because of its very chatty, Positive conversational tone that I found it a little bit hard sometimes to think critically about the content I wanted to allow it to kind of wave over me and just Accept this a very kind of positive kind of friendly dynamic between these two characters and so I found it easier to then look at the written transcript, in order to try to Figure out what was being said and if there are any kind of mistakes or if anything was misleading So there is something very superficially attractive and very plausible about the way in which the content is presented within the podcast.

[00:20:22] **Christoph Bartneck:** Minyi, how did you experience the podcast?

[00:20:25] **Minyi:** Actually for me, the podcast is, I don't have too much experience for that, but I think it's a good way to learn something easily and And if you have a rest or you're in relaxed, you can, you know, to learn something very coarse.

[00:20:42] **Christoph Bartneck:** Well, when I listened to it, I had a similar, but maybe still slightly different experience than this. I was wondering whether The anthropomorphization of the language models does this to us. The more humanlike it becomes, the more it plays into our what's the term for it?

Suspension of disbelief, is the one, right? Where we just apply humanlevel Competence onto whatever is being produced. Now think about it, if you go back to ChatGPT, the words appear one after the other. It looks like it's typing. It doesn't have to be this way, right? The whole sentence could just come up as one.

And just the impression like, Oh, it is thinking. It is saying something. It is speaking. That makes a lot of difference. And then now we've got language models that speak even before this podcast thing. And now we have a podcast or a language model that has Dialogue which makes it even more human like

[00:21:52] **Michael-John:** and even in a way The kind of friendliest most positive human you could imagine the dialogue between the two characters was just extraordinarily positive. You don't want to disagree with these voices There's something almost hyperhuman about it in terms of the persuasiveness, because an ordinary person Human being is going to sound a little grumpy and disagreeable it at times. And so I think that plays even more into the desire to believe or suspend disbelief. Maybe.

[00:22:20] **Christoph Bartneck:** So is this a good method for learning?

[00:22:25] **Minyi:** Actually I would like to see the content that generated by AI is really amazing. It will be a good form of learning. But not now, because the content that is generated is not accurate for me. And I think the beginner are unable to distinguish whether the content is right or wrong, or good or bad, since the lack of the necessary knowledge about that. So, perhaps we will get the better outcomes after changing the LLM for many times, I think.

[00:23:05] **Michael-John:** I agree, it could be a useful tool for learning and like any other learning tool, we have to learn how to learn. We have to learn how to use it and gain those kinds of critical thinking skills which enable us to evaluate it now.

Of course they're improving all the time. I think in many ways it did provide a pretty reasonable and engaging synopsis, despite the reservations and criticisms which I've made about some of the content. As a naive consumer of the content, if you expect that this will replace reading or engaging with the original research, then I think there could be some quite serious dangers to that.

As a tool to help us supplement learning to maybe provide a quick and accessible synopsis, especially as the technology improves. I can imagine that there would be some benefits to it.

[00:23:56] **Christoph Bartneck:** So using ChattGPT or any language models for the purpose of a language tool like a spell checker or so forth. We're coming to terms with this by now. But isn't there a danger if you trust language models too much and they only know the superficial, they only know the banalities. They only know what seems to be common knowledge, but it could be terribly wrong.

[00:24:23] **Michael-John:** As a, as an ethicist, I think this raises very interesting questions because large language models largely reflect back to us. They are as someone like Shannon Vala says, a kind of mirror the what's already there in the training data.

Now, of course, the data that they're trained on be biased in certain ways, may reflect historical patterns of injustice, for example. And then if we look to them for advice or guidance in the ethical domain, then we might entrench past patterns of ethical injustice. Injustice rather than doing the hard and at this point anyway, I think human work of imagining alternative possibilities of asking ourselves, could we live in somewhat different ways than we might see reflected in the training data that the language model is using, [00:25:14] **Christoph Bartneck:** which brings us then to the question of what is the purpose of the podcast? If it is to quickly summarize something on a superficial level, what It might work, but if you want it to be precise, or specific, or detailed, or critical, or disagreeable, it is likely to fail.

[00:25:36] **Michael-John:** Well, it struck me that there's something almost approaching a paradox, maybe not a strict paradox here. But I was thinking about this question of would I use it in my research. And the answer is that I wouldn't want to use it unless I was able to understand the original research myself in order to check for its reliability and so on.

But if I can understand the original research and engage with it, Then why would I need the synopsis provided by, the language model or the podcast in this case? On the other hand, if I couldn't understand the original research and I'd be in the position of a rather naive consumer of whatever It is that the language model is producing, and quite susceptible to the introduction of biases and so on.

[00:26:19] **Minyi:** Can I ask a question to you, Christopher? Sure. Absolutely. Yeah. So, I would like to you mentioned two maybe the functions of the podcast. So, what do you think of our podcast? What's the function of our podcast?

[00:26:32] **Christoph Bartneck:** You mean the Human Robot Interaction podcast?

[00:26:33] **Minyi:** Yes.

Hmm, that's a difficult one. Well, clearly this is not a thesis, not a PhD thesis, so barely ever anything new comes of it. Is it informative? I, no, there's some novelty in it. Sometimes I report on interesting things that were not commonly known. So there's an, a definite aspect of novelty to it, but then again, there are other podcasts.

So if you look at stuff you should know, which is a very good podcast that introduces all sorts of topics and interests in history and society, technology, and it provides a good summary of it and I love to listen to it to better understand things around me, let's say so. and there's a place for this.

And I'll probably put the language model podcast in this category for that could work. For what we are doing is, well that's where it gets interesting. Cause now we're having a critical discussion. We are potentially even having some new thoughts, or at least we're offering something new. That large language model couldn't. So we have to justify our existence here.

[00:27:42] **Michael-John:** I mean, this question to me is a very interesting one concerning the possibility of novelty or creativity. So I described large language models as like mirrors. But of course, they're not perfectly reflecting back what's already been done. There are nice examples of generative AI which can produce music, so Suno for example.

And you can ask it to produce music on a theme and given a particular genre. Now, what it doesn't do at the moment is create music. Interesting new genres, but of course we know from the history of music that people who we think of as innovative We're standing on the shoulders of Giants So I imagine that there would be no Elvis Presley without blues And I imagine there'll be no Beatles without Elvis Presley and so on and so forth So we can ask well, what is it that we as humans are doing so very different when we think of ourselves as being creative and so on if not in some sense building on what's gone before and offering new patterns and associations. So I don't know the answers to these questions, but I think we've got something we can ask. Minjy.

[00:28:49] **Christoph Bartneck:** So you are right now tasked with doing something new.

How do you do it?

[00:29:03] **Minyi:** In general, for most PhD students, if we created something new or we do something new, we just to read the paper first and, oh, just the same like AI. We upload some papers to AI and, we need to search some papers for us and to read them, to select the important things or critical things that what we would like, are we interested in that to research?

And I would like to see this. Really interesting, because we can process the information from the paper to know what is useful for my research. But I don't know how the AI to select this content, because I'm pretty curious about the content selection in LLM, even in the podcast that we generated.

It's just really interesting. And okay, back to the I just want to point back to the question, because after the review of the paper, there's a literature review, and we will come up with the. Research questions and try to based on the paper that what we find to, I would like to do some experiment to impractically to address some phenomenons. For example, the ethical symmetry that we're researching now. So

[00:30:28] **Christoph Bartneck:** so when you sit down. And you have to think about, what is my next experiment going to be? How do you decide on what you're going to do?

[00:30:42] **Minyi:** That's a good question for me, because In my research if we're speaking of the novelty, I would like to say that, I don't know how to explain that, because it's, I think it's the cognitive process in my brain, and to think, firstly, we reviewed all the papers we find, and to set a gap, and what they what they didn't do before, and then to think, okay, they didn't do that. And if it is meaningful, if it is significant, and if it is, maybe we can to do that.

[00:31:27] **Christoph Bartneck:** So the trick is to know what you don't know.

[00:31:32] **Minyi:** Something like that, yeah.

[00:31:34] **Christoph Bartneck:** Which is, I think, the classical definition of novelty, originality, isn't it?

[00:31:39] **Michael-John:** Well, this is what Socrates once said, that he was told by the Delphic Oracle that he was the wisest man in Greece. And he thought, well, this can't be true. People around me seem to know. And so he went and he questioned people, reputed knowledgeable by others, and he says tongue in cheek, especially by themselves. And came to the conclusion that yes, indeed, he was the wisest man in all of Greece because he alone knew what he didn't know. And so this is This is quite a, quite a challenge to understand our own limitations.

I find the question of creativity very interesting as well. I think, Christophe, you asked a very difficult question. How is it that we generate new ideas ourselves? And I find that I spend a lot of time thinking about a question, reading widely, making notes, and then sometimes ideas just come to me when I'm on a walk or in the shower or, Some people report having ideas coming to them in dreams.

I have to say that all the ideas that come to me in dreams are dreadful. But other people have more luck in that regard than I do. And some of those ideas seem to make sense when I think about them further and others seem to be no good at all. So I think there is a mystery about the kind of basis of idea generation you mentioned.

And I think you're right that there's a question about the opacity of large language models. So they're like black boxes. It's hard to see what's going on in the inside even if you understand the basic kind of architecture of them which generates problems in its own right. And I think there's a similar kind of opacity or black box quality to human creativity at large. What about you, Christoph? What's your process like for generating novel ideas?

[00:33:20] **Christoph Bartneck:** To one degree, it is diversity of inputs. So I expose myself to fields of knowledge or things that have nothing to do with what I do otherwise. And sometimes I see these bridges where I go like, hey, wait a minute. Over here they use this in that way.

Could I use this in my context? So that's one thing is combining different fields of knowledge. And the other thing is similar to what you just said. I just have to sit on it. So I have to get to the depth of it and then be with it without holding on too tight and just letting it go. And for me personally, when I go swimming, I, my mind wanders, which is much to the distress of everybody else in my lane because I can't keep on counting the right distances. But my mind wanders. And when the mind wanders, new ideas happen and indeed I have no idea how.

[00:34:24] **Michael-John:** This is a slight tangent I'm sure but I think there's something to be said in favor of boredom and one thing that we all have with smartphones in our pockets or smartwatches on our wrist is an excuse never to be bored. But sometimes allowing the mind to wander as we're swimming or walking in the hills or whatever it might be allows us to somehow or another generate interesting ideas.

[00:34:45] **Minyi:** Yeah, the true I would like to say there are some scenarios that make me, inspiration and to get some new idea. The first scenario that I would like to mention is the, because I prefer to work in the night and when I'm in the office and only myself, I just walk around and to think about things and we can get an idea.

And the second thing is also about the swimming, because I really like swimming and I think It's a totally quiet place when you dip into the water. So you can to think about what the plan, the next plan, and I will, in, in my mind, I will to review all the papers that I read and to find what's the importance and what's the significance, yeah, something like that.

[00:35:32] **Christoph Bartneck:** All right. So the overall challenge remains for large language models. to be creative. They might have some abilities to summarize, but only on a superficial level. So coming back to the really, really hard question, should we actually continue doing podcasts?

[00:35:55] **Michael-John:** Are you worried, Christophe, that your job as a podcast host is in jeopardy?

[00:36:01] **Christoph Bartneck:** To some degree, yes. If I make a podcast which is just a summary of something, then yes. I hope that I would have more to offer.

[00:36:14] **Minyi:** Okay, for me I think it depends. Because if we think the podcaster can teach the, some audience some things. scientist things or the academic things. So I think we should do it continue because the AI has a long term to work. I think it's the long way to training them to be more mature. But If just for fun, but I think to comparing to doing the podcast, I think we should to training the AI generate LLM.

[00:36:54] **Michael-John:** I think that the resurgence in interest in podcasts over recent years is so wonderful and surprising because it's effectively, you know, asynchronous radio. And I think that one of the issues that we'll face going forward, not just in the. Podcast space, but elsewhere is how to deal with this enormous influx of information so we can generate Podcasts very quickly very cheaply very easily by the millions And then the role of the podcast the podcast host becomes more a gatekeeping Function to point the listener towards podcasts that may be more or less interesting, and to make these discriminations and also to find people, hopefully, with relevant expertise, or someone who can say something interesting about the topic.So there's a more curatorial, I think, function to the podcast host.

[00:37:46] **Christoph Bartneck:** All right. Thank you so much for this interesting conversation and I hope this podcast continues.

[00:37:54] **Michael-John:** Long may, long may you last as a podcaster.

[00:37:56] **Minyi:** Bye.